UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Statistical Models for Genetics

2526-1-F8205B022

Obiettivi formativi

l corso si propone di fornire agli studenti gli strumenti statistici per analizzare e interpretare i dati genetici, con particolare riferimento agli studi di associazione su scala genomica (GWAS). Al termine del corso lo studente sarà in grado di leggere criticamente un GWAS pubblicato, riconoscendone il controllo di qualità, il modello statistico e la strategia di replicazione; di eseguire in autonomia, in ambiente R, una pipeline completa che va dal controllo di qualità dei genotipi al test di associazione, fino alla correzione per la struttura di popolazione, all'imputazione e alla meta-analisi; di scegliere modelli e codifiche del genotipo appropriati alla domanda biologica; di diagnosticare e correggere le insidie più comuni di questi studi, in primo luogo la stratificazione di popolazione e la molteplicità dei test; e infine di tradurre un segnale statistico in un'ipotesi biologica preliminare, producendo un report riproducibile che documenti l'intera analisi.

Contenuti sintetici

Il corso copre l'intero percorso analitico di uno studio di associazione genetica, organizzato attorno a un unico caso di studio — un GWAS caso-controllo per il diabete di tipo 2 in popolazione europea. I temi trattati comprendono il vocabolario di base della variazione genetica e i formati dei dati, le frequenze alleliche, l'equilibrio di Hardy–Weinberg e il linkage disequilibrium; la natura del dato di sequenziamento e la lettura critica della sua qualità; il controllo di qualità a livello di campione e di variante; la correzione della struttura di popolazione mediante analisi delle componenti principali e modelli misti; il test di associazione single-SNP con i relativi problemi di interpretazione dell'effetto e di molteplicità dei test; l'imputazione genotipica; la meta-analisi e la replicazione; e infine l'annotazione funzionale che porta dal locus statistico all'ipotesi biologica.

Programma esteso

Il corso si apre con un'introduzione alla variazione genetica, in cui si fissano i concetti e il vocabolario fondamentali (locus, allele, genotipo, polimorfismo, SNP, aplotipo) e si discute la natura osservazionale della genetica umana, che rende indispensabile il ragionamento statistico.

Si passa quindi ai dati genetici veri e propri: i formati con cui si rappresentano i genotipi, il calcolo delle frequenze alleliche e genotipiche, l'equilibrio di Hardy–Weinberg come modello di riferimento e il suo test, e il linkage disequilibrium con le sue implicazioni per il disegno e l'interpretazione di un GWAS.

Un modulo integrativo è dedicato a come nasce il dato di sequenziamento, seguendo la catena che porta dalla molecola di DNA al file di sequenze. L'obiettivo non è bioinformatico ma statistico: comprendere che il dato è una stima rumorosa e imparare a leggere criticamente un report di qualità, riconoscendo le sorgenti di rumore strutturato che si propagano ai modelli a valle.

Il controllo di qualità viene poi affrontato come parte integrante di ogni GWAS, distinguendo i filtri a livello di campione da quelli a livello di variante e discutendo il razionale statistico di ciascuna soglia, fino alla documentazione del processo.

Il corso entra quindi nel tema della struttura di popolazione, mostrando come la stratificazione non corretta gonfi i falsi positivi e come la si possa diagnosticare e correggere, attraverso l'analisi delle componenti principali e i modelli lineari misti, che affrontano in un solo passo sia la stratificazione sia le parentele criptiche.

Il nucleo del corso è il test di associazione single-SNP: il quadro della regressione lineare e logistica, le diverse codifiche del genotipo, l'interpretazione dell'effetto in termini di effect size e odds ratio, il problema della molteplicità dei test e l'origine della soglia genome-wide, e la diagnostica complessiva dell'analisi (Manhattan plot, QQ plot, fattore di inflazione genomica).

Seguono l'imputazione genotipica, che estende la copertura dai marcatori del chip all'intero genoma sfruttando il linkage disequilibrium, e il lavoro con i dosaggi nei test di associazione; e la meta-analisi con la replicazione, che permette di combinare più coorti per guadagnare potenza, valutarne l'eterogeneità e consolidare i risultati.

Il corso si chiude con la fase post-GWAS, in cui il segnale statistico viene collegato alla biologia tramite annotazione funzionale e risorse pubbliche, e con un progetto guidato su un secondo dataset, in cui gli studenti replicano l'intera pipeline e producono un report in stile clinico.

Prerequisiti

Si richiede una solida preparazione in biostatistica classica: inferenza statistica, test d'ipotesi, intervalli di confidenza e familiarità con i modelli di regressione lineare e logistica. È necessaria una conoscenza di base del linguaggio R e dell'ambiente RStudio, sufficiente a manipolare dati e produrre grafici. Non è richiesta alcuna competenza pregressa di genetica o di bioinformatica: i concetti biologici necessari vengono introdotti nel corso nella misura utile all'analisi statistica.

Metodi didattici

Il corso alterna lezioni frontali e laboratori pratici in R, condotti sui portatili degli studenti in ambiente RStudio, senza necessità di infrastrutture di calcolo dedicate. L'impianto didattico è guidato da un unico caso di studio percorso dall'inizio alla fine: ogni nuovo metodo viene introdotto a partire dalla domanda clinica o epidemiologica che lo motiva, sviluppato come modello statistico e infine tradotto in codice. I laboratori non sono esercitazioni isolate ma costruiscono progressivamente i diversi segmenti di una pipeline unica, che confluisce nell'elaborato finale. La

lettura guidata di articoli scientifici accompagna le lezioni, per abituare lo studente a interpretare criticamente metodi, statistica e figure di un GWAS pubblicato.

Modalità di verifica dell'apprendimento

La verifica si basa su un elaborato finale costituito da un report riproducibile in R Markdown, costruito su un secondo dataset caso-controllo diverso da quello usato a lezione e fornito dal docente. L'esame si articola in tre componenti tra loro collegate: un progetto guidato, svolto in laboratorio supervisionato, in cui lo studente replica l'intera pipeline e dà forma all'elaborato; la consegna del report, vincolata al requisito che esso sia rieseguibile integralmente sul portatile del docente, dal dato grezzo alle tabelle e alle figure, senza interventi manuali; e una discussione orale di circa quindici minuti, individuale o di piccolo gruppo, in cui si presentano e si difendono criticamente le scelte metodologiche e i risultati.

L'elaborato e la sua discussione sono valutati rispetto a quattro criteri di pari peso, ciascuno pari al venticinque per cento del giudizio complessivo. Il primo è la correttezza metodologica, cioè l'appropriatezza di ogni passo della pipeline rispetto alla domanda e ai dati, con giustificazione esplicita delle scelte di codifica, covariate e soglie. Il secondo è l'interpretazione statistica, ossia la capacità di leggere i risultati in termini di effect size, confidenza e incertezza, e non soltanto di p-value. Il terzo è l'interpretazione clinico-biologica, vale a dire la collocazione dei loci principali nel loro contesto biologico mediante risorse pubbliche, con la dovuta cautela sulla distinzione tra locus e gene. Il quarto è la riproducibilità, intesa come capacità del report di essere rieseguito end-to-end senza interventi manuali.

Il voto è espresso in trentesimi. La sufficienza, fissata a diciotto trentesimi, richiede una pipeline corretta nei passaggi essenziali e un report che si riesegua senza errori; il punteggio pieno richiede inoltre un'interpretazione statistica e clinica accurata e scelte metodologiche pienamente giustificate. La lode è riservata agli elaborati che mostrano autonomia critica, ad esempio attraverso analisi di sensibilità non richieste, la discussione delle assunzioni violate o un uso particolarmente lucido delle risorse di annotazione. La riproducibilità costituisce un requisito di ammissione alla valutazione: un report che non viene eseguito è restituito per la correzione prima di essere valutato nel merito. Il lavoro in piccolo gruppo è ammesso per il progetto, ma la discussione orale verifica il contributo e la comprensione individuali. La frequenza dei laboratori, pur non valutata di per sé, è fortemente raccomandata, poiché ciascuno di essi costruisce una parte della pipeline che confluisce nell'elaborato.

Testi di riferimento

Per una panoramica concisa dell'intera pipeline GWAS si rimanda a Bush e Moore, Genome-Wide Association Studies (PLOS Computational Biology, 2012), e al tutorial applicativo di Marees e colleghi, A tutorial on conducting genome-wide association studies (Int J Methods Psychiatr Res, 2018). Per gli approfondimenti tematici si consigliano Price e colleghi, New approaches to population stratification in genome-wide association studies (Nature Reviews Genetics, 2010); Marchini e Howie, Genotype imputation for genome-wide association studies (Nature Reviews Genetics, 2010); ed Evangelou e Ioannidis, Meta-analysis methods for genome-wide association studies and beyond (Nature Reviews Genetics, 2013). Come sintesi conclusiva è particolarmente utile Visscher e colleghi, 10 years of GWAS discovery: biology, function, and translation (American Journal of Human Genetics, 2017). Per il modulo sul dato di sequenziamento si rimanda a Goodwin, McPherson e McCombie, Coming of age: ten years of next-generation sequencing technologies (Nature Reviews Genetics, 2016), e, sull'origine del punteggio di qualità, a Ewing e Green, Base-calling of automated sequencer traces using phred. II (Genome Research, 1998).

Periodo di erogazione dell'insegnamento

secondo semestre

Lingua di insegnamento

italiano

Sustainable Development Goals

SALUTE E BENESSERE